L'INDICE DEI LIBRI DEL MESE
Mensile
Data 06-2019
Pagina 12
Foglio 1 / 3

Approcci sistemici alla gestione dell'informazione

Grandi dati e grossi problemi

di Davide Lovisolo

La comparsa, pochi decenni fa, di internet e della connettività globale ha dato origine a un fenomeno assolutamente nuovo: un accumulo di enormi quantità di dati conservati in banche digitali, la cui quantità raddoppia ogni pochi giorni e in prospettiva ogni poche ore. Tutto è dati: immagini, grafici, foto, informazioni sulle preferenze, azioni e posizione di miliardi di esseri umani. E tutti questi dati possono essere analizzati, elaborati e se ne può estrarre informazione, tutta con potenziale valore economico; è a questa realtà che si dà il nome di Big Data (BD), intendendo sia la mole dei dati in sé che la possibilità di utilizzarla con procedure sempre più automatiche e veloci (gli algoritmi, del cui potere e dei cui rischi ha scritto Gabriele Lolli sull'"Indice" di dicembre 2018). Cominciano ad uscire anche in Italia libri che affrontano specificamente il tema dei BD, a volte con un approccio problematico (Domenico Talia, *La società calcolabile e i big data. Algoritmi e persone nel mondo digitale,* pp. 224, € 13, Rubbettino Soveria Mannelli CZ 2018), in altri casi con un taglio più entusiastico. Un esempio è il libro di Viktor Mayer-Schönberger e Thomas Ramge, *Reinventare il capitalismo nell'era dei big data* (trad. dall'inglese di Giuseppe Maugeri, pp. 212, € 24, Egea, Milano 2018), che, pur presentando gli aspetti critici del fenomeno e avanzando alcune proposte interessanti, vede le nuove tecnologie fondate sui dati come capaci di rifondare il capitalismo su basi nuove. In generale, questi approcci partono sempre dal ruolo dei BD nei campi dell'economia e del lavoro. Ma se è vero che tutti i dati hanno un valore economico, l'accumulo di informazione e del loro trattamento ha anche risvolti in un certo senso più rilevanti sulla formazione di conoscenza: è la questione affrontata nel libro di Sabina Leonelli, *La ricerca scientifica nell'era dei Big Data. Cinque modi in cui i Big Data danneggiano la scienza e come salvarla.* (Meltemi 2018) che merita un'attenzione particolare proprio perché affronta un aspetto poco discusso: il ruolo dei BD nella ricerca scientifica. L'autrice insegna filosofia e storia della scienza all'Università di Exeter, è consulente della Commissione europea, e oltre a decine di articoli ha pubblicato nel 2016 il libro *Data-Centric Biology: a Philosophical Study* (Chicago University Press, 2016); ora esce questo contributo in italiano, il primo che affronti questo tema. Data la formazione dell'autrice, l'approccio è di tipo solidamente epistemologico. Si parte dalle aspettative, che sono grandi in tutti i settori, ed anche in particolare in campo scientifico: la disponibilità dei BD e la possibilità di mettere in relazione dati di origine diversa può svelare correlazioni finora non evidenziabili, con conseguenze anche traslazionali, come la possibilità di monitorare la diffusione di malattie infettive o di prevenire disastri ambientali. I Big Data sono sovente associati ad un'altra grande promessa: gli Open Data, i dati della scienza aperta. Associare alla enorme mole di dati e alla loro velocità di generazione l'idea di una loro totale accessibilità è l'aspetto veramente rivoluzionario. In questo quadro, i dati cambiano status: da oggetto privato del ricercatore, che lui solo può interpretare correttamente e selezionando quali rendere pubblici (approccio teoria-centrico), si passa ora ad una visione datocentrica: i dati sono "entità pubbliche che hanno valore scientifico indipendentemente dal loro ruolo di prova per una determinata ipotesi e che possono essere interpretati in modi diversi a seconda (...) degli interessi dei ricercatori che li analizzano". Questo comporta una "rivalutazione radicale del potenziale dei dati di generare conoscenza". In realtà, la promessa degli Open Data (OD) è molto difficile da realizzare, anche per i vincoli privatistici e l'estrema competitività del settore; il loro potenziale valore economico non sempre ne favorisce l'apertura, ma piuttosto la compravendita. La discussione si polarizza sovente su posizioni estreme, anche se, come spiega Leonelli, che di questo specifico aspetto è una riconosciuta autorità, ci sono tentativi di ragionevole compromesso. Una proprietà fondamentale dei BD è la loro mobilità, la capacità di viaggiare ed essere riusati in situazioni diverse. Senza questa proprietà i BD non avrebbero senso; proprio per questo è importante poter disporre di dati formattati con procedure standard: questo può essere più facile ad esempio per i dati ricavati da GPS; ma i dati biologici e medici in genere non sono così standardizzati. Tale "pluralismo scientifico" non è casuale, ma è dovuto al fatto che dati di questo tipo sono ottenuti con metodologie e tecniche specializzate alla descrizione di specifiche proprietà degli oggetti e dei fenomeni studiati. Gli archivi non possono essere contenitori di dati grezzi, ma devono essere esplicitati i criteri di ordinamento, le procedure da utilizzare per richiamarli. Il problema è che i potenziali utilizzatori sono in genere persone che non hanno informazione sulle condizioni in cui sono stati prodotti. E qui l'autrice pone un primo pesante mattone teorico: in questa luce, il problema è manageriale ma anche e soprattutto epistemologico; "l'organizzazione dei dati è condizionata dal contesto e dallo scopo per cui vengono analizzati". Il problema non è risolvibile con soluzioni universali o completamente automatizzate: richiede giudizi informati. Gli archivi devono poter essere interoperabili, cioè i dati devono poter essere utilizzati da esperti in campi diversi. Questa interoperabilità comporta che i dati viaggino da un sito ad un altro; nel percorso si trasformano nella forma e nel contenuto, ed è errato pensarli come rappresentazioni immutabili e fedeli della realtà. Leonelli si è quindi proposta il compito di seguire i viaggi dei dati da un archivio all'altro: compito difficile, perché sovente le tracce non sono ben documentate. Solo una piccola parte delle banche dati, le migliori, hanno una visione chiara ed esplicita della loro gestione e della loro evoluzione nel tempo; fare questo comporta grandi investimenti per la manutenzione, controlli e coinvolgimento dell'utenza. È errato pensare, come fanno molti governi che tendono a tagliare le spese per la ricerca, che i BD siano una via più economica per accelerare il progresso scientifico. Il rischio è di fare danni, e il secondo capitolo elenca i cinque rischi principali. *Il*

conservativismo: se le banche non vengono aggiornate, si tende a mantenere categorie prestabilite e si sviluppa un atteggiamento dogmatico. È comune trovare confusione e ignoranza sui criteri usati nel corso degli anni, soprattutto in banche con più di 10 anni. Continuare a usare dati vecchi, di archivi sostanzialmente defunti, può provocare seri danni, in particolare nelle scienze della vita, dove si ha a che fare con oggetti che evolvono anche rapidamente. *I dati inattendibili*: i controlli di qualità e i criteri per decidere quali dati sono buoni richiedono tempo e risorse. È difficile avere standard universali. È necessario decontestualizzare (per rendere i dati universalmente utilizzabili e affidabili), e poi ricontestualizzare (per verificarne la validità rispetto all'applicazione specifica). *I dati parziali*: ignorare la natura selettiva (insita nel processo di produzione di conoscenza scientifica) dei dati che si trovano nelle banche porta a mistificare la realtà. Molti dati (fotografie, immagini cliniche) sono difficilmente analizzabili senza strumenti complessi, e questo genera tra l'altro gerarchie fra siti potenti e quelli tecnologicamente più deboli, come molti siti africani. Una conseguenza è che molti dati biomedici si riferiscono in prevalenza ad appartenenti a fasce medio-alte dei paesi sviluppati, in particolare a bianchi e maschi, generando informazione fuorviante, che rischia potenzialmente di aumentare le diseguaglianze: il "digital divide" produce "data divide". *La corruzione dei dati*: nonostante gli sforzi e l'impegno degli enti pubblici, la stragrande maggioranza dei dati a scopo di ricerca è generata in ambito commerciale e privato. La filosofia dei grandi operatori, come Google, è chiara: i dati personali, in quanto facilmente accessibili, sono pubblici, ma nello stesso tempo privatizzabili e vendibili. Il caso di Cambridge Analytica insegna. *I dati sensibili*: conseguenza del punto precedente. La disponibilità di enormi quantità di informazioni contiene la potenzialità di comprendere meglio le esigenze dei cittadini, ma comporta seri rischi. L'autrice racconta il proprio coinvolgimento in un progetto mirato a confrontare dati medici, climatici e messaggi di Twitter per capire quanto contino le condizioni climatiche sulle malattie respiratorie stagionali. L'idea era buona, ma si è scontrata con vari problemi, primo di tutti l'uso di Twitter, che dà i dati (o almeno una parte) per scopi scientifici, mentre altri operatori lo fanno solo a pagamento; ma Twitter è utilizzato principalmente da una fascia giovane e urbana, e fornisce quindi una rappresentazione limitata e distorta. Sarebbero necessarie ulteriori ricerche per filtrare i dati, ma richiedono tempo e risorse aggiuntive, scontrandosi con il fatto che i governi vogliono risposte chiare senza tanti *caveat* sui limiti, e i ricercatori sono sotto pressione per pubblicare. Infine, c'è il problema del duplice uso: i dati da Twitter possono essere usati per migliorare la vita della gente, ma anche per sorvegliarla. Conclusione: le questioni etiche devono essere viste come parte integrante delle esigenze tecniche e scientifiche associate alla gestione e all'analisi dei dati, e non come estranee. Per poter affrontare questi rischi, è necessario per Leonelli tornare all'approccio epistemologico. Cosa sono i dati? Per la visione prevalente, sono una rappresentazione sistematica e ripetibile della realtà, dotata di contenuto fisso e immutabile. A differenza dell'informazione che ricaviamo dai nostri sensi, non dipendono dalla percezione di un unico individuo. I BD da questo punto di vista sono una grande opportunità: più fatti si accumulano e si legano fra loro, più conoscenza se ne può derivare. È quella che l'autrice chiama visione rappresentativa dei dati. Ma i dati non parlano mai da soli. Il modo in cui vengono analizzati e interpretati dipende almeno in parte dalle conoscenze di chi li studia. È proprio la possibilità di usare metodi e criteri diversi che rende i BD e OD così produttivi e rivoluzionari: un esempio interessante e positivo è la cosiddetta *citizen science*, la produzione di dati da parte di osservatori amatoriali, poi utilizzati da ricercatori che ne cambiano formato e contenuti per confrontarli con dati ottenuti per altre vie, analizzandoli statisticamente, normalizzandoli, e così via. I dati viaggiano, e come lo fanno conta. Tutto ciò non è facilmente conciliabile con la visione rappresentativa, a cui la Leonelli oppone la visione relazionale: il significato assegnato ai dati non dipende solo dalle caratteristiche fisiche, ma "anche dalle motivazioni e dagli strumenti usati (...) e l'affidabilità(...) dal rigore dei processi usati per produrli e analizzarli". Questa visione, che incoraggia la cura dei dati e della loro storia, può andare incontro a problemi e obiezioni, che l'autrice analizza con chiarezza. Ne citerò qui solo una: non c'è il rischio di cadere nel relativismo? La risposta è no, anzi questo approccio "incoraggia alla continua vigilanza su motivi, sui metodi e sugli scopi per cui certi oggetti vengono (...) proposti come dati a supporto di determinate asserzioni". Incoraggia anche il confronto fra fonti diverse: la conferma di un'interpretazione è più credibile se deriva da dati che non condividono la stessa storia. Viene qui accennato, un po' di sfuggita, il problema della riproducibilità dei dati, considerato classicamente uno dei pilastri del metodo scientifico, ma oggi oggetto di riflessione critica anche all'interno della comunità scientifica (si veda l'editoriale *Repetitive flaws* sul numero del 20 gennaio 2016 di "Nature": https://www.nature.com/news/repetitive-flaws-1.19192). Su queste basi, nella parte finale del libro vengono avanzate alcune proposte per "incoraggiare il meglio". Anche qui, il cardine è il legame fra scienza ed etica; è necessario un confronto continuo con i gruppi sociali in grado di contribuire alla valutazione delle scelte: proprio in quanto i BD abbattono le separazioni rigide fra competenze diverse, devono anche servire per mettere in comunicazione settori sociali diversi. Concezione della produzione scientifica che non esclude il mondo esterno, ma lo coinvolge. Le scelte non sono automatiche: l'idea che chi fa algoritmi per Google non sa come verranno usati rischia di spingerci verso un determinismo tecnologico. Il rischio è che le tecnologie vengano considerate enti superiori all'uomo. Nelle scelte deve entrare la valutazione delle potenziali conseguenze dell'innovazione. Nel caso dell'intelligenza artificiale, quali caratteristiche privilegiare nello sviluppo di nuovi algoritmi? Quale riorganizzazione sociale e culturale è necessaria? Come si potrebbe modificare la ricerca per produrre meno discriminazione? Trovare risposte non è facile, ma non per questo si deve arrestare l'innovazione: l'acquisizione di conoscenza porta sempre vantaggi e svantaggi, e il processo va governato. Un esempio di come la cattiva gestione si ritorca contro chi l'ha praticata è il caso di Google Flu Trends, che utilizzava ricerche su Google per creare un predittore di epidemie di influenza, basandosi sul fatto che la gente cerca "influenza" anche prima o invece di andare dal

dottore. Il problema è che non si era fatta un'analisi seria sulle motivazioni di chi fa la ricerca, sui termini usati, ecc. Il risultato, al di là delle uscite trionfalistiche, è stato un fallimento: nel 2013 non riuscì a predire un'epidemia di grandi dimensioni, e nel 2015 segnalò un numero dei casi doppio di quelli effettivamente verificatisi. Un'analisi approfondita di questa vicenda, diventata esempio di come si possa creare conoscenza inattendibile, si trova in un lavoro di ricercatori di Harvard, *The Parable of Google Flu: Traps in Big Data Analysis* (http://nrs.harvard.edu/urn-3:HUL.InstRepos:12016836). Coinvolgimento sociale non vuol dire quindi che tutti devono essere esperti informatici, ma che ci devono essere forme di comunicazione fra tecnici ed utenti: non è scambio simmetrico, ma bisogna provare. E il processo non va idealizzato: ci sono fasi in cui i ricercatori operano in maniera separata, e questo è utile per elaborare soluzioni che vadano oltre la contingenza. Nelle conclusioni vorrei solo ricordare due proposte: quella di un codice di comportamento analogo a quello dei medici, e quella della "Slow Science", rallentare per riflettere.

davide.lovisolo@unito.it

D. Lovisolo ha insegnato biologia all'Università di Torino